Quantifying transmission dynamics of acute hepatitis C virus infections in a heterogeneous population using sequence data

Gonché Danesh[1,*],Victor Virlogeux[2],Christophe Ramière[3],Caroline Charre[3], Laurent Cotte[4‡], Samuel Alizon[1‡]

1 MIVEGEC (UMR CNRS 5290, IRD, UM), Montpellier, France

2 Clinical Research Center, Croix-Rousse Hospital, Hospices Civils de Lyon, France

3 Virology Laboratory, Croix-Rousse Hospital, Hospices Civils de Lyon, France

4 Infectious Diseases Department, Croix-Rousse Hospital, Hospices Civils de Lyon, France

‡These authors contributed equally to this work.

* Corresponding author: gonche.danesh@ird.fr

Abstract

Opioid substitution and syringes exchange programs have drastically reduced hepatitis C virus (HCV) spread in France but HCV sexual transmission in men having sex with men (MSM) has recently arisen as a significant public health concern. The fact that the virus is transmitting in a heterogeneous population, with 'new' and 'classical' hosts, makes prevalence and incidence rates poorly informative. However, additional insights can be gained by analyzing virus phylogenies inferred from dated genetic sequence data. By combining a phylodynamics approach based on Approximate Bayesian Computation (ABC) and an original transmission model, we estimate key epidemiological parameters of an ongoing HCV epidemic among MSMs in Lyon (France). We show that this new epidemic is largely independent of the 'classical' HCV epidemics and that its doubling time is ten times lower (0.44 years *versus* 4.37 years). These results have practical implications for HCV control and illustrate the additional information provided by virus genomics in public health.

Background

It is estimated that 71 million people worldwide suffer from chronic hepatitis C virus (HCV) infections [1, 2]. The World Health Organisation (WHO) and several countries have issued recommendations towards the 'elimination' of this virus, which they define as an 80% reduction in new chronic infections and a 65% decline in liver mortality by 2030 [2]. HIV-HCV coinfected patients are targeted with priority because of the shared transmission routes between the two viruses [3] and because of the increased virulence of HCV in coinfections [4–6]. Successful harm reduction interventions, such as needle-syringe exchange and opiate substitution programs, as well as a high level of enrolment into care programs for HIV-infected patients, have led to a drastic drop in the prevalence of active HCV infections in HIV-HCV coinfected patients in several European countries during the recent years [7–10]. Unfortunately, this elimination goal is challenged by the emergence of HCV sexual transmission, especially among men having sex with men (MSM). This trend is reported to be driven by unprotected sex, drug use in the context of sex ('chemsex'), and potentially traumatic practices such as fisting [11–13]. The epidemiology of HCV infection in the Dat'AIDS cohort has been extensively described from 2000 to 2016 [14–16]. The incidence of acute HCV infection has been estimated among HIV-infected MSM between 2012 and 2016, among HIV-negative MSM enrolled in PrEP between in 2016-2017 [13] and among HIV-infected and HIV-negative MSMs from 2014 to 2017 [17]. In the area of Lyon (France), HCV incidence has been shown to increase concomitantly with a shift in the profile of infected hosts [17]. Understanding and quantifying this recent increase is the main goal of this study.

Several modelling studies have highlighted the difficulty to control the spread of HCV infections in HIV-infected MSMs in the absence of harm reduction interventions [12, 18]. Furthermore, we recently described the spread of HCV from HIV-infected to HIV-negative MSMs, using HIV pre-exposure prophylaxis (PrEP) or not, through shared high-risk practices [17]. More generally, an alarming incidence of acute HCV infections in both HIV-infected and PrEP-using MSMs was reported in France in 2016-2017 [13]. Additionally, while PrEP-using MSMs are regularly screened for HCV, those who are HIV-negative and do not use PrEP may remain undiagnosed and untreated for years. In general, we know little about the population size and practices of HIV-negative MSM who do not use PrEP. All these epidemiological events could jeopardize the goal of HCV elimination by creating a large pool of infected and undiagnosed patients, which could fuel new infections in intersecting populations. Furthermore, the epidemiological dynamics of HCV infection have mostly been studied in intravenous drug users (IDU) [19–22] and the general population [23, 24]. Results from these populations are not easily transferable to other populations, which calls for a better understanding of the epidemiological characteristics of HCV sexual transmission in MSM.

Given the lack of knowledge about the focal population driving the increase in HCV incidence, we analyse virus sequence data with phylodynamics methods. This research field has been blooming over the last decade and hypothesizes that the way rapidly evolving viruses spread leaves 'footprints' in their genomes [25–27]. By combining mathematical modelling, statistical analyses and phylogenies of infections, where each leaf corresponds to the virus sequence isolated from a patient, current methods can infer key parameters of viral epidemics. This framework has been successfully applied to

other HCV epidemics [28–31], but the ongoing one in Lyon is challenging to analyze because the focal population is heterogeneous, with 'classical' hosts (typically HIV-negative patients infected through nosocomial transmission or with a history of opioid intravenous drug use or blood transfusion) and 'new' hosts (both HIV-infected and HIV-negative MSM, detected during or shortly after acute HCV infection phase, potentially using recreational drugs such as cocaine or cathinones), where host profiles have been established by field epidemiologists based on interviews and risk factors. Our phylodynamics analysis relies on an Approximate Bayesian Computation (ABC, [32]) framework that was recently developed and validated using a simple Susceptible-Infected-Recovered (SIR) model [33].

Assuming an epidemiological transmission model with two host types, 'classical' and 'new' (see the Methods), we use dated virus sequences to estimate the date of onset of the HCV epidemics in 'classical' and 'new' hosts, the level of mixing between hosts types, and, for each host type, the duration of the infectious period and the effective reproduction ratio (i.e. the number of secondary infections, [34]). To validate our results we performed a parametric bootstrap analysis, we tested the sensitivity of the method to differences in sampling proportions between the two types of hosts. We also tested the sensitivity of the method to phylogenetic reconstruction uncertainty, and we performed a cross-validation analysis to explore the robustness of our inference framework. We find that the doubling time of the epidemics is one order of magnitude lower in 'new' than in 'classical' hosts, therefore emphasising the urgent need for public health action.

Results

The phylogeny inferred from the dated virus sequences shows that 'new' hosts (in red) tend to be grouped in clades (Figure 1). This pattern suggests a high degree of assortativity in the epidemics (i.e. hosts tends to infect hosts from the same type). The ABC phylodynamics approach allows us to go beyond a visual description and to quantify several epidemiological parameters.

As for any Bayesian inference method, we need to assume a prior distribution for each parameter. These priors, shown in grey in Figure 2, are voluntarily designed to be large and uniformly distributed to be as little informative as possible. One exception is the date of onset of the epidemics, for which we use the output of the phylogenetic analysis conducted in Beast (see the Methods) as a prior. We also assume the date of the 'new' hosts epidemics to be after 1997 based on epidemiological data.

The inference method converges towards posterior distributions for each parameter, which are shown in red in Figure 2. The estimate for the origin of the epidemic in 'classical' hosts is $t_0 = 1957.47$ $[1948.61; 1961.96]$ (numbers in brackets indicate the 95% Highest Posterior Density, or HPD). For the 'new' host type, we were not able to estimate when the epidemic (t_2) has started.

We find the level of assortativity between host types to be high for 'classical' ($a_1 = 0.94$ $[0.83; 1.0]$) as well as for 'new' hosts ($a_2 = 0.92$ $[0.81; 0.99]$). Therefore, hosts mainly infect hosts from the same type.

The phylodynamics approach also allows us to infer the duration of the infectious period for each host type. Assuming that this parameter does not vary over time, we estimate it to be 3.85 years $[1.09; 8.33]$ for 'classical' hosts (parameter $1/\gamma_1$) and 0.45 years $[0.30; 0.77]$ for 'new' hosts

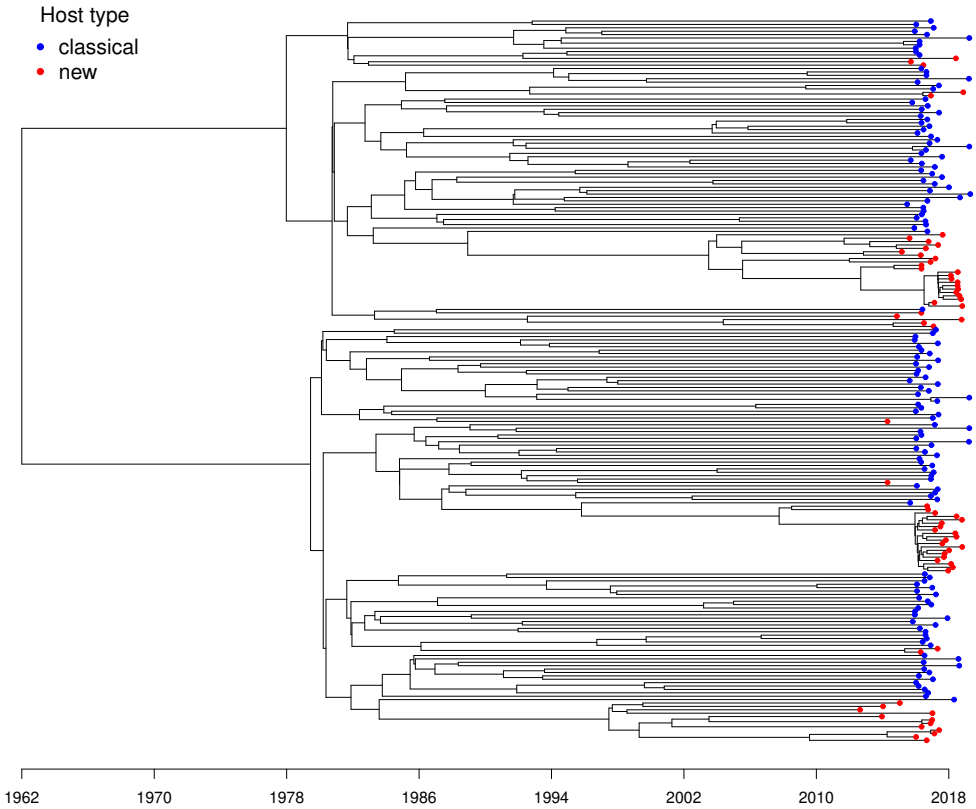

Fig 1. Phylogeny of HCV infections in the area of Lyon (France). 'Classical' hosts are in blue and 'new' hosts are in red. Sampling events correspond to the end of black branches. The phylogeny was estimated using Bayesian inference (Beast2). See the Methods for additional details.

(parameter $1/\gamma_2$). We compute the ratio of γ_2/γ_1 and the 95% credibility interval does exclude 1.

Regarding effective reproduction numbers, i.e. the number of secondary infections caused by a given host over its infectious period, we estimate that of 'classical' hosts to have decreased from $R^{(1),t_1} = 1.96$ $[1.45; 3.29]$ to $R^{(1),t_2} = 1.61$ $[1.05; 2.08]$ after the introduction of the third-generation HCV test in 1997. The inference on the differential transmission parameter indicates that HCV transmission rate is $\nu = 9.0$ $[7.7; 9.9]$ times greater from 'new' hosts than from 'classical' hosts. By combining these results (see the Methods), we compute the effective reproduction number in 'new' hosts and find $R^{(2),t_3} = 1.73$ $[1.03; 4.32]$. We compute the ratio of the $R(t)$ of 'new' hosts over the $R(t)$ of 'classical' hosts after 1997 and, the median value is 1.14 and the 95% credibility interval is $[0.56; 3.25]$.

To better understand the differences between the two host types, we compute the epidemic doubling time (t_D), which is the time for an infected population to double in size. t_D is computed for each type of host, assuming complete assortativity (see the Methods). We find that for the 'classical'

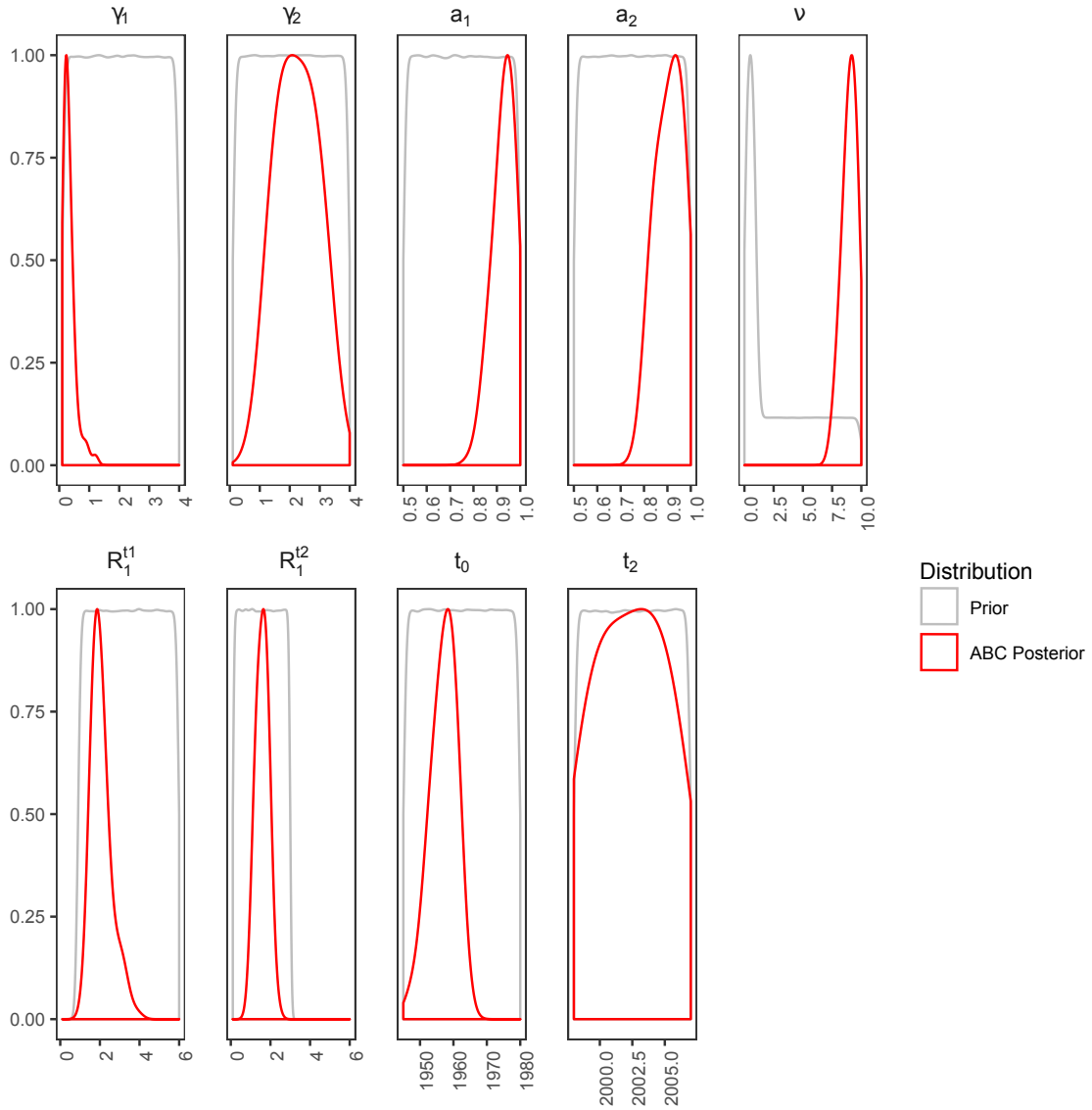

Fig 2. Parameter prior and posterior distributions. Prior distributions are in grey and posterior distributions inferred by ABC are in red. The thinner the posterior distribution width, the more accurate the inference. Posterior distributions are truncated based on the prior distribution.

hosts, before 1997 $t_D^{(1),t1} \approx 2.8$ years ([1.1; 5.0] years). After 1997, the pace decreases with a doubling time of $t_D^{(1),t2} \approx 4.4$ years ([2.0; 20.8] years). For the epidemics in the 'new' hosts, we estimate that $t_D^{(2),t3} \approx 0.44$ years ([0.09; 8.84] years). When computing the ratio of the doubling times of classical hosts after 1997 over the doubling times of the new hosts ($t_D^{(1),t2}/t_D^{(2),t3}$) to estimate the current difference we find that $t_D^{(1),t2}$ is 10 times higher than $t_D^{(2),t3}$ with a 95% credibility interval of [0.62; 14.99]. However, the 75% credibility interval does exclude 1 and is [3.39; 25.61]. Distributions for theses three doubling times are shown in Supplementary Figure S2.

Supplementary Figure S3 shows the correlations between parameters based on the posterior distributions. We mainly find that the R_t in 'classical' hosts after the introduction of the third

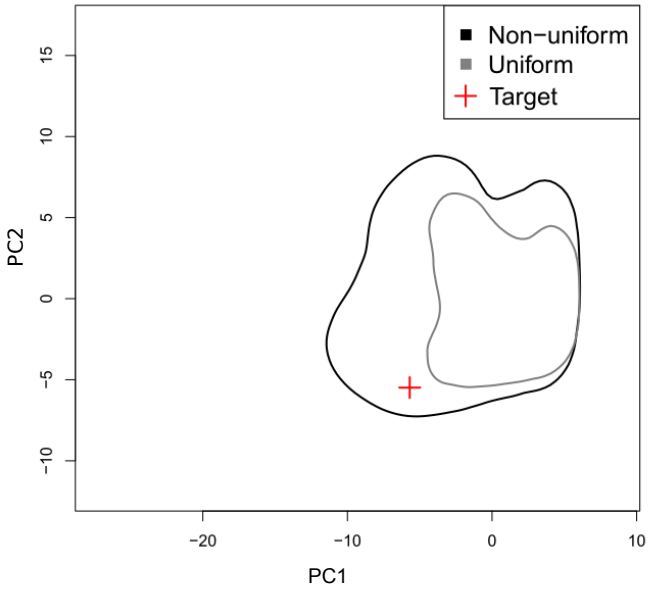

Fig 3. Goodness-of-fit estimated using parameter bootstrap. The graph displays envelopes containing 90% of the 10,000 simulations for each distribution. The envelope in black results from the posterior distribution, in grey, results from the uniform distribution drawn from the 95% HPD distribution. The target data is represented by a red cross. Axes units are based on the outcome of principal component analysis using the simulated summary statistics.

generation of HCV detection tests (i.e. $R^{(1),t_2}$) is negatively correlated to ν and positively correlated to γ_2. In other words, if the epidemic spreads rapidly in 'classical' hosts, it requires a slower spread in 'new' hosts to explain the phylogeny. $R_0^{(1),t_2}$ is also slightly negatively correlated to γ_1, which most likely comes from the fact that for a given R_0, epidemics with a longer infection duration have a lower doubling time and therefore a weaker epidemiological impact. Overall, these correlations do not affect our main results, especially the pronounced difference in infection periods (γ_1 and γ_2).

To validate these results, we performed a goodness-of-fit by simulating phylogenies using the resulting posterior distributions to determine whether these are similar to the target dataset (see the Methods). In Figure 3, we see that the target data in red, i.e. the projection of the observed summary statistics from the phylogeny shown in Figure 1, is contained in the envelope containing 90% of the simulations drawn from the posterior distributions. If we use the 95% HPD of the posterior but assume a uniform distribution instead of the true posterior distribution, we find that the target phylogeny is not contained in the envelope. These results confirm that the posterior distributions we infer are highly informative.

To further explore the robustness of our inference method, we use simulated data to perform a 'leave one out' cross-validation (see the Methods). As shown in Supplementary Figure S5, the mean relative error made for each parameter inference is limited and comparable to what was found using a simpler SIR model [33]. One exception is for the 'new' hosts' level of assortativity (a_2). This is

likely due to the poor signal given the small size of the observed phylogeny.

A potential issue is that the sampling rate of 'new' hosts may be higher than that of 'classical' hosts. To explore the effect of such sampling biases on the accuracy of our results, we sub-sampled the 'new' hosts population by pruning the target phylogeny, i.e. randomly removing 50% of the 'new' hosts' tips. In Supplementary Figure **??** we show the posterior distributions estimated by our ABC method using the different pruned phylogenies. We find that although the confidence intervals are wider, the posterior distributions are all similar with the posterior distributions estimated using the target phylogeny. Finally, to evaluate the impact of phylogenetic reconstruction uncertainty, we analysed 100 additional trees from the Beast posterior distribution. In Supplementary figure **??**, we show that the estimates from our ABC method are qualitatively similar for all these trees.

Discussion

Over the last years, the area of Lyon (France) witnessed an increase in HCV incidence both in HIV-positive and HIV-negative populations of men having sex with men (MSM) [17]. This increase appears to be driven by sexual transmission and echoes similar trends in Amsterdam [35] and Switzerland [36]. A quantitative analysis of the epidemic is necessary to optimise public health interventions. Unfortunately, this is challenging because the monitoring of the population at risk is limited and because classical tools in quantitative epidemiology, especially incidence time series, are poorly informative with such a heterogeneous population. To circumvent this problem, we used HCV sequence data, which we analysed using phylodynamics. To account for host heterogeneity, we extended and validated an existing Approximate Bayesian Computation framework [33].

From a public health point of view, our results have two major implications. First, we find a strong degree of assortativity in both 'classical' and 'new' host populations. The virus phylogeny does hint at this result (Figure 1) but the ABC approach allows us to quantify the pattern and to show that assortativity may be higher for 'classical' hosts. The second main result has to do with the striking difference in doubling times. Indeed, the current spread of the epidemics in 'new' hosts appears to be five times more rapid than the spread in the 'classical' hosts in the early 1990s before the advent of the third generation tests in 1997, and ten times more rapid that the spread in the 'new' hosts after 1997. That the duration of the infectious period in 'new' hosts is in the same order of magnitude as the time until treatment suggests that the majority of the transmission events may be occurring during the acute phase. This underlines the necessity to act rapidly upon detection, for instance by emphasising the importance of protection measures such as condom use and by initiating treatment even during the acute phase [37]. A better understanding of the underlying contact networks could provide additional information regarding the structure of the epidemics and, with that respect, next-generation sequence (NGS) data could be particularly informative [38–40].

Some potential limitations of the study are related to the sampling scheme, the assessment of the host type, and the transmission model. Regarding the sampling, the proportion of infected 'new' host that is sampled is unknown but could be high. For the 'classical' hosts, we selected a representative subset of the patients detected in the area but this sampling is likely to be low. However, the effect of underestimating sampling for the new epidemics would be to underestimate its

spread, which is already faster than the classical epidemics. When running the analyses on different phylogenies with half of the 'new' hosts sequences, we find results similar to those obtained with the whole phylogeny, suggesting that our ABC framework is partly robust to sampling biases. In general, implementing a more realistic sampling scheme in the model would be possible but it would require a more detailed model and more data to avoid identifiability issues. Regarding assigning hosts to one of the two types, this was performed by clinicians independently of the sequence data. The main criterion used was the infection stage (acute or chronic), which was complemented by other epidemiological criteria (history of intravenous drug use, blood transfusion, HIV status). Finally, the 'classical' and the 'new' epidemics appear to be spreading on contact networks with different structures. However, such differences are beyond the level of details of the birth-death model we use here and would require a larger dataset for them to be inferred.

To test whether the infection stage (acute vs. chronic) can explain the data better than the existence of two host types, we developed an alternative model where all infected hosts first go through an acute phase before recovering or progressing to the chronic phase. As for the model with two host types, we used three time intervals. Supplementary Figure **??** shows the diagram of the model as well as the corresponding equations. Interestingly, it was almost impossible to simulate phylogenies with this model, most likely because of its intrinsic constrains on assortativity (both acute and chronic infections always generate new acute infections).

To our knowledge, few attempts have been made in phylodynamics to tackle the issue of host population heterogeneity. In 2018, a study used the structured coalescent model to investigate the importance of accounting for so-called 'superspreaders' in the recent Ebola epidemics in West Africa [41]. The same year, another study used the birth-death model to study the effect of drug resistance mutations on the R_0 of HIV strains [42]. Both of these are implemented in Beast2. Although the multi-type birth-death model is unlikely to be directly applicable to this HCV epidemic because it links the two epidemics via mutation (a host of type A becomes a host of type B), whereas here the links is done via transmission events (a host of type A infects a host of type B), we ran an analysis using the BEAST 2 package bdmm with our data. We were unable to conclude anything from this analysis which rises the limitation of the likelihood-based approach for this dataset.

Overall, we show that our ABC approach, which we validated for simple SIR epidemiological models [33], can be applied to more elaborate models that current phylodynamics methods have difficulties to capture. Further increasing the level of details in the model may require to increase the number of simulations but also to introduce new summary statistics. Another promising perspective would be to combine sequence and incidence data. Although this could not be done here due to the limited sampling, such data integration can readily be done with regression-ABC.

Material and methods

Epidemiological data

The Dat'AIDS cohort is a collaborative network of 23 French HIV treatment centres covering approximately 25% of HIV-infected patients followed in France (Clinicaltrials.gov ref NCT02898987).

Host profiles have been established by field epidemiologists based on interviews and risk factors.

HCV sequence data

We included HCV molecular sequences of all MSM patients diagnosed with acute HCV genotype 1a infection at the Infectious Disease Department of the Hospices Civils de Lyon, France, and for whom NS5B sequencing was performed between January 2014 and December 2017 ($N = 68$). HCV genotype 1a isolated from $N = 145$ non-MSM, HIV-negative, male patients of similar age were analysed by NS5B sequencing at the same time for phylogenetic analysis. This study was conducted following French ethics regulations. All patients gave their written informed consent to allow the use of their personal clinical data. The study was approved by the Ethics Committee of Hospices Civils de Lyon.

HCV testing and sequencing

HCV RNA was detected and quantified using the Abbott RealTime HCV assay (Abbott Molecular, Rungis, France). The NS5B fragment of HCV was amplified between nucleotides 8256 and 8644 by RT-PCR as previously described and sequenced using the Sanger method. Electrophoresis and data collection were performed on a GenomeLab^{TM} GeXP Genetic Analyzer (Beckman Coulter). Consensus sequences were assembled and analysed using the GenomeLab^{TM} sequence analysis software. The genotype of each sample was determined by comparing its sequence with HCV reference sequences obtained from GenBank.

Nucleotide accession numbers

All HCV NS5B sequences isolated in MSM and non-MSM patients reported in this study were submitted to the GenBank database. The list of Genbank accession numbers for all sequences is provided in Appendix.

Dated viral phylogeny

To infer the time-scaled viral phylogeny from the alignment we used a Bayesian Skyline model in BEAST v2.5.2 [43]. The general time-reversible (GTR) nucleotide substitution model was used with a strict clock rate fixed at $1.3 \cdot 10^{-3}$ based on data from Ref. [44] and a gamma distribution with four substitution rate categories. The MCMC was run for 100 million iterations and samples were saved every 100,000 iterations. We selected the maximum clade credibility using TreeAnnotator BEAST2 package. The date of the last common ancestor was estimated to be 1961.95 with a 95% Highest Posterior Density (HPD) of $[1941.846; 1975.516]$.

Epidemiological model and simulations

We assume a Birth-Death model with two hosts types (Supplementary Figure S1) with 'classical' hosts (numbered 1) and new hosts (numbered 2). This model is described by the following system of ordinary differential equations (ODEs):

Table 1. Prior distributions for the birth-death model parameters over the three time intervals. t_0 is the date of origin of the epidemics in the studied area, t_1 is the date of introduction of 3rd generation HCV tests, t_2 is the date of emergence of the epidemic in 'new' hosts and t_3 is the time of the most recent sampled sequence.

Interval	γ_i	ν	$R^{(1)}$	a_i
$[t_0, t_1]$	Unif$(0.1, 4)$	0	Unif$(0.9, 6)$	Unif$(0.5, 1)$
$[t_1, t_2]$			Unif$(0.1, 3)$	
$[t_2, t_3]$		Unif$(0, 1)$ & Unif$(1, 10)$		

$$\frac{dI_1}{dt} = a_1\beta I_1 + (1 - a_2)\nu\beta I_2 - \gamma_1 I_1 \tag{1a}$$

$$\frac{dI_2}{dt} = a_2\beta\nu I_2 + (1 - a_1)\beta I_1 - \gamma_2 I_2 \tag{1b}$$

In the model, transmission events are possible within each type of hosts and between the two types of hosts at a transmission rate β. Parameter ν corresponds to the transmission rate differential between classical and new hosts. Individuals can be 'removed' at a rate γ_1 from an infectious compartment (I_1 or I_2) via infection clearance, host death or change in host behaviour (e.g. condom use). The assortativity between host types, which can be seen as the percentage of transmissions that occur with hosts from the same type, is captured by parameter a_i.

The effective reproduction number (denoted R_t) is the number of secondary cases caused by an infectious individual in a fully susceptible host population [34]. We seek to infer the R_t from the classical epidemic, denoted $R^{(1)}$ and defined by $R^{(1)} = \beta/\gamma_1$, as well as the R_t of the new epidemic, denoted $R^{(2)}$ and defined by $R^{(2)} = \nu\beta/\gamma_2 = \nu R^{(1)}\gamma_1/\gamma_2$.

The doubling time of an epidemic (t_D) corresponds to the time required for the number of infected hosts to double in size. It is usually estimated in the early stage of an epidemic when epidemic growth can assumed to be exponential. To calculate it, we assume perfect assortativity ($a_1 = a_2 = 1$) and approximate the initial exponential growth rate by $\beta - \gamma_1$ for 'classical' hosts and $\nu\beta - \gamma_2$ for 'new' hosts. Following [45], we obtain $t_D^{(1)} = \ln(2)/(\beta - \gamma_1)$ and $t_D^{(2)} = \ln(2)/(\nu\beta - \gamma_2)$.

We consider three time intervals. During the first interval $[t_0, t_1]$, t_0 being the year of the origin of the epidemic in the area of Lyon, we assume that only classical hosts are present. The second interval $[t_1, t_2]$, begins in $t_1 = 1997.3$ with the introduction of the third generation HCV tests, which we assume to have affected $R^{(1)}$ through the decrease of the transmission rate β. Finally, the 'new' hosts appear during the last interval $[t_2, t_3]$, where t_2, which we infer, is the date of origin of the second outbreak. The final time (t_3) is set by the most recent sampling date in our dataset (2018.39). The prior distributions used are summarized in Table 1 and shown in Figure 2. Given the phylogeny structure suggesting a high degree of assortativity, we assume the assortativity parameters, a_1 and a_2, to be higher than 50%. For the prior distribution of parameter ν, we combined a uniform distribution from 0 to 1 with a uniform distribution from 1 to 10. This was done to ensure that the probability to have $\nu < 1$ is equal to the probability to have $\nu > 1$.

To simulate phylogenies, we use our TiPS simulator [46] implemented in R via the Rcpp package. This is done in a two-step procedure. First, epidemiological trajectories are simulated using the compartmental model in equation 1 and Gillespie's stochastic event-driven simulation algorithm [47]. The number of individuals in each compartment and the reactions occurring through the simulations of trajectories, such as recovery or transmission events, are recorded. Using the target phylogeny, we know when sampling events occur. For each simulation, each sampling date is randomly associated to a host compartment using the observed fraction of each infection type (here 68% of the dates associated with 'classical' hosts type and 32% with 'new' hosts). Once the sampling dates are added to the trajectories, we move to the second step, which involves simulating the phylogeny. This step starts from the last sampling date and follows the epidemiological trajectory through a coalescent process, that is backwards-in-time. Each backward step in the trajectory can induce a tree modification given a probability and the population size: a sampling event leads to a labelled leaf in the phylogeny, a transmission event can lead to the coalescence of two sampled lineages or to no modification of the phylogeny (if one of the lineages is not sampled).

We implicitly assume that the sampling rate is low, which is consistent with the limited number of sequences in the dataset. We also assume that the virus can still be transmitted after sampling.

We simulate $60,000$ phylogenies from known parameter sets drawn in the prior distributions shown in Table 1. These are used to perform the rejection step and build the regression model in the Approximate Bayesian Computation (ABC) inference.

ABC inference

Summary statistics

Phylogenies are rich objects and to compare them we break them into summary statistics. These are chosen to capture the epidemiological information of interest. In particular, following an earlier study, we use summary statistics from branch lengths, tree topology, and lineage-through-time (LTT) [33], and summary statistics based on the Laplacian spectrum using the spectR function of the RPANDA R package [48].

We also compute new summary statistics to extract information regarding the heterogeneity of the population, the assortativity, and the difference between the two R. To do so, we annotate each internal node by associating it with a probability to be in a particular state (here the host type, 'classical' or 'new'). We assume that this probability is given by the ratio

$$P(Y) = \frac{\text{number of descendent leaves labelled } Y}{\text{number of descendent leaves}} \tag{2}$$

where Y is a state (or host type). Each node is therefore annotated with n ratios, n being the number of possible states. Since in our case $n = 2$, we only follow one of the labels and use the mean and the variance of the distribution of the ratios (one for each node) as summary statistics.

In a phylogeny, cherries are pairs of leaves that are adjacent to a common ancestor. There are $n(n+1)/2$ categories of cherries. Here, we compute the proportion of homogeneous cherries for each label and the proportion of heterogeneous cherries. We also consider pitchforks, which we define as a cherry and a leaf adjacent to a common ancestor, and introduce three categories: homogeneous

pitchforks, pitchforks whose cherries are homogeneous for a label and whose leaf is labelled with another trait, and pitchforks whose cherries are heterogeneous.

The Lineage-Through-Time (LTT) plot displays the number of lineages of a phylogeny over time. In this plot, the number of lineages is incremented by one every time there is a new branch in the phylogeny and is decreased by one every time there is a new leaf in the phylogeny. We use the ratios defined for each internal node to build an LTT plot for each label type, which we refer to as 'LTT label plot'. After each branching event in phylogeny, we increment the number of lineages by the value of the ratio of the internal node for the given label. This number of lineages is decreased by one every time there is a leaf in the phylogeny. In the end, we obtain $n = 2$ LTT label plots.

Finally, for each label, we compute some of our branch lengths summary statistics on homogeneous clades and heterogeneous clades present in the phylogeny. Homogeneous clades are defined by their root having a ratio of 1 for one type of label and their size being greater than $N_{\min}$. For heterogeneous clades, we keep the size criterion and impose that the ratio is smaller than 1 but greater than a threshold ϵ. After preliminary analyses, we set $N_{\min} = 4$ leaves and $\epsilon = 0.7$. We then obtain a set of homogeneous clades and a set of heterogeneous clades, the branch lengths of which we pool into two sets to compute the summary statistics of heterogeneous and homogeneous clades. Note that we always select the largest clade, for both homogeneous and heterogeneous cases, to avoid redundancy.

Regression-ABC

We first measure multicollinearity between summary statistics using variance inflation factors (VIF). Each summary statistic is kept if its VIF value is lower than 10. This stepwise VIF test leads to the selection of 101 summary statistics out of 330.

We then use the `abc` function from the `abc` R package [49] to infer posterior distributions generated using only the rejection step. Finally, we perform linear adjustment using an elastic net regression.

The `abc` function performs a classical one-step rejection algorithm [32] using a tolerance parameter P_δ, which represents a percentile of the simulations that are close to the target. To compute the distance between a simulation and the target, we use the Euclidian distance between normalized simulated vectors of summary statistics and the normalized target vector.

Before linear adjustment, the `abc` function performs smooth weighting using an Epanechnikov kernel [32]. Then, using the glmnet package in R, we implement an elastic-net (EN) adjustment, which balances the Ridge and the LASSO regression penalties [50]. Since the EN performs a linear regression, it is not subject to the risk of over-fitting that may occur for non-linear regressions (e.g. when using neural networks, support vector machines or random forests).

In the end, we obtain posterior distributions for t_0, t_2, a_1, a_2, ν, γ_1, γ_2, $R^{(1),t_1}$ and $R^{(1),t_2}$ using our ABC-EN regression model with $P_\delta = 0.05$.

Parametric bootstrap and cross-validation

Our goodness-of-fit validation consists in simulating $10,000$ additional phylogenies from parameter sets drawn in posterior distributions. We then compute summary statistics and perform a goodness

of fit using the gfitpca function from the abc R package [49]. The function performs principal component analysis (PCA) using the simulated summary statistics. It displays envelopes containing a given percentage, here 90%, of the simulations. The projection of the observed summary statistics is displayed to check if they are contained or not in the envelopes. If the posterior distribution is informative, we expect the target data to be contained in the envelope. This analysis was performed either on the posterior distribution, or on a uniform distribution based on the 95% HPD posterior distribution of each parameter, the latter being less informative.

To assess the robustness of our ABC-EN method to infer epidemiological parameters of our BD model, we also perform a 'leave-one-out' cross-validation as in [33]. This consists in inferring posterior distributions of the parameters from one simulated phylogeny, assumed to be the target phylogeny, using the ABC-EN method with the remaining $59,999$ simulated phylogenies. We run the cross-validation 100 times with 100 different target phylogenies. We consider three parameter distributions θ: the prior distribution, the prior distribution reduced by the feasibility of the simulations and the ABC inferred posterior distribution. For each of these parameter distributions, we measure the median and compute, for each simulation scenario, the mean relative error (MRE) such as:

$$MRE = \frac{1}{100} \sum_{i=1}^{100} \mid \frac{\theta_i}{\Theta} - 1 \mid \tag{3}$$

where Θ is the true value.

Acknowledgements

We thank Jūlija Pečerska for her help with Beast2. GD is funded by the Fondation pour la Recherche Médicale (FRM grant number ECO20170637560). GD and SA acknowledge further support from the CNRS, the IRD and the itrop HPC (South Green Platform) at IRD Montpellier, which provided HPC resources that contributed to the results reported here (`https://bioinfo.ird.fr/`).

References

[1] Messina JP, Humphreys I, Flaxman A, Brown A, Cooke GS, Pybus OG, et al. Global distribution and prevalence of hepatitis C virus genotypes. Hepatology. 2015;p. 77–87.

[2] European Union HCV Collaborators. Hepatitis C virus prevalence and level of intervention required to achieve the WHO targets for elimination in the European Union by 2030: a modelling study. Lancet Gastroenterol Hepatol. 2017;2(5):325–336.

[3] Alter MJ. Epidemiology of viral hepatitis and HIV co-infection. J Hepatol. 2006;44(S1):S6–9.

[4] Rosenthal E, Salmon-Céron D, Lewden C, Bouteloup V, Pialoux G, Bonnet F, et al. Liver-related deaths in HIV-infected patients between 1995 and 2005 in the French GERMIVIC Joint Study Group Network (Mortavic 2005 Study in collaboration with the Mortalité 2005 survey, ANRS EN19)*. HIV Medicine. 2009;10(5):282–289.

[5] Kovari H, Ledergerber B, Cavassini M, Ambrosioni J, Bregenzer A, Stöckle M, et al. High hepatic and extrahepatic mortality and low treatment uptake in HCV-coinfected persons in the Swiss HIV cohort study between 2001 and 2013. Journal of Hepatology. 2015 Sep;63(3):573–580.

[6] Klein MB, Althoff KN, Jing Y, Lau B, Kitahata M, Lo Re V, et al. Risk of End-Stage Liver Disease in HIV-Viral Hepatitis Coinfected Persons in North America From the Early to Modern Antiretroviral Therapy Eras. Clin Infect Dis. 2016 Nov;63(9):1160–1167.

[7] Pradat P, Pugliese P, Poizot-Martin I, Valantin MA, Cuzin L, Reynes J, et al. Direct-acting antiviral treatment against hepatitis C virus infection in HIV-Infected patients – "En route for eradication"? Journal of Infection. 2017 Sep;75(3):234–241. Available from: `http://www.sciencedirect.com/science/article/pii/S0163445317301421`.

[8] Béguelin C, Suter A, Bernasconi E, Fehr J, Kovari H, Bucher HC, et al. Trends in HCV treatment uptake, efficacy and impact on liver fibrosis in the Swiss HIV Cohort Study. Liver International. 2018;38(3):424–431.

[9] Berenguer J, Jarrín I, Pérez-Latorre L, Hontañón V, Vivancos MJ, Navarro J, et al. Human Immunodeficiency Virus/Hepatits C Virus Coinfection in Spain: Elimination Is Feasible, but the Burden of Residual Cirrhosis Will Be Significant. Open Forum Infect Dis. 2018 Jan;5(1).

[10] Boerekamps A, van den Berk GE, Lauw FN, Leyten EM, van Kasteren ME, van Eeden A, et al. Declining Hepatitis C Virus (HCV) Incidence in Dutch Human Immunodeficiency Virus-Positive Men Who Have Sex With Men After Unrestricted Access to HCV Therapy. Clin Infect Dis. 2018 Apr;66(9):1360–1365.

[11] van de Laar T, Pybus O, Bruisten S, Brown D, Nelson M, Bhagani S, et al. Evidence of a Large, International Network of HCV Transmission in HIV-Positive Men Who Have Sex With Men. Gastroenterology. 2009 May;136(5):1609–1617.

[12] Salazar-Vizcaya L, Kouyos RD, Zahnd C, Wandeler G, Battegay M, Darling KEA, et al. Hepatitis C virus transmission among human immunodeficiency virus-infected men who have sex with men: Modeling the effect of behavioral and treatment interventions. Hepatology. 2016;64(6):1856–1869.

[13] Pradat P, Huleux T, Raffi F, Delobel P, Valantin MA, Poizot-Martin I, et al. Incidence of new hepatitis C virus infection is still increasing in French MSM living with HIV. AIDS. 2018 May;32(8):1077.

[14] Pradat P, Caillat-Vallet E, Sahajian F, Bailly F, Excler G, Sepetjan M, et al. Prevalence of hepatitis C infection among general practice patients in the Lyon area, France. Eur J Epidemiol. 2001 Jan;17(1):47–51.

[15] D'Oliveira JA, Voirin N, Allard R, Peyramond D, Chidiac C, Touraine JL, et al. Prevalence and sexual risk of hepatitis C virus infection when human immunodeficiency virus was acquired through sexual intercourse among patients of the Lyon University Hospitals, France, 1992-2002. J Viral Hepat. 2005 May;12(3):330–332. Available from: `http://europepmc.org/abstract/med/15850476`.

[16] Sahajian F, Bailly F, Vanhems P, Fantino B, Vannier-Nitenberg C, Fabry J, et al. A randomized trial of viral hepatitis prevention among underprivileged people in the Lyon area of France. J Public Health (Oxf). 2011 Jun;33(2):182–192.

[17] Ramière C, Charre C, Miailhes P, Bailly F, Radenne S, Uhres AC, et al. Patterns of Hepatitis C Virus Transmission in Human Immunodeficiency Virus (HIV)–infected and HIV-negative Men Who Have Sex With Men. Clin Infect Dis. 2019;.

[18] Virlogeux V, Zoulim F, Pugliese P, Poizot-Martin I, Valantin MA, Cuzin L, et al. Modeling HIV-HCV coinfection epidemiology in the direct-acting antiviral era: the road to elimination. BMC Medicine. 2017 Dec;15(1):217.

[19] Pybus OG, Cochrane A, Holmes EC, Simmonds P. The hepatitis C virus epidemic among injecting drug users. Infection, Genetics and Evolution. 2005 Mar;5(2):131–139.

[20] Sweeting MJ, De Angelis D, Hickman M, Ades AE. Estimating hepatitis C prevalence in England and Wales by synthesizing evidence from multiple data sources. Assessing data conflict and model fit. Biostatistics. 2008 Oct;9(4):715–734.

[21] Kwon JA, Iversen J, Maher L, Law MG, Wilson DP. The Impact of Needle and Syringe Programs on HIV and HCV Transmissions in Injecting Drug Users in Australia: A Model-Based Analysis. JAIDS Journal of Acquired Immune Deficiency Syndromes. 2009 Aug;51(4):462.

[22] Pitcher AB, Borquez A, Skaathun B, Martin NK. Mathematical modeling of hepatitis c virus (HCV) prevention among people who inject drugs: A review of the literature and insights for elimination strategies. Journal of Theoretical Biology. 2018 Nov;.

[23] Breban R, Arafa N, Leroy S, Mostafa A, Bakr I, Tondeur L, et al. Effect of preventive and curative interventions on hepatitis C virus transmission in Egypt (ANRS 1211): a modelling study. The Lancet Global Health. 2014 Sep;2(9):e541–e549.

[24] Heffernan A, Cooke GS, Nayagam S, Thursz M, Hallett TB. Scaling up prevention and treatment towards the elimination of hepatitis C: a global mathematical model. The Lancet. 2019 Mar;393(10178):1319–1329.

[25] Grenfell BT, Pybus OG, Gog JR, Wood JL, Daly JM, Mumford JA, et al. Unifying the epidemiological and evolutionary dynamics of pathogens. Science. 2004;303(5656):327–32.

[26] Volz EM, Koelle K, Bedford T. Viral phylodynamics. PLoS Comput Biol. 2013;9(3):e1002947.

[27] Frost SD, Pybus OG, Gog JR, Viboud C, Bonhoeffer S, Bedford T. Eight challenges in phylodynamic inference. Epidemics. 2015;10:88–92.

[28] Pybus OG, Charleston MA, Gupta S, Rambaut A, Holmes EC, Harvey PH. The epidemic behavior of the hepatitis C virus. Science. 2001;292(5525):2323–5.

[29] Magiorkinis G, Magiorkinis E, Paraskevis D, Ho SYW, Shapiro B, Pybus OG, et al. The Global Spread of Hepatitis C Virus 1a and 1b: A Phylodynamic and Phylogeographic Analysis. PLOS Medicine. 2009 Dec;6(12):e1000198.

[30] Stadler T, Kühnert D, Bonhoeffer S, Drummond AJ. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proc Natl Acad Sci USA. 2013;110(1):228–33.

[31] Joy JB, McCloskey RM, Nguyen T, Liang RH, Khudyakov Y, Olmstead A, et al. The spread of hepatitis C virus genotype 1a in North America: a retrospective phylogenetic study. The Lancet Infectious Diseases. 2016 Jun;16(6):698–702.

[32] Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian Computation in Population Genetics. Genetics. 2002 Dec;162(4):2025–2035.

[33] Saulnier E, Gascuel O, Alizon S. Inferring epidemiological parameters from phylogenies using regression-ABC: A comparative study. PLOS Computational Biology. 2017 Mar;13(3):e1005416.

[34] Anderson RM, May RM. Infectious Diseases of Humans. Dynamics and Control. Oxford: Oxford University Press; 1991.

[35] van de Laar TJW, van der Bij AK, Prins M, Bruisten SM, Brinkman K, Ruys TA, et al. Increase in HCV Incidence among Men Who Have Sex with Men in Amsterdam Most Likely Caused by Sexual Transmission. J Infect Dis. 2007 Jul;196(2):230–238.

[36] Wandeler G, Gsponer T, Bregenzer A, Günthard HF, Clerc O, Calmy A, et al. Hepatitis C Virus Infections in the Swiss HIV Cohort Study: A Rapidly Evolving Epidemic. Clin Infect Dis. 2012 Nov;55(10):1408–1416.

[37] AASLD/IDSA HCV Guidance Panel. Hepatitis C guidance: AASLD-IDSA recommendations for testing, managing, and treating adults infected with hepatitis C virus. Hepatology. 2015 Sep;62(3):932–954.

[38] Romero-Severson EO, Bulla I, Leitner T. Phylogenetically resolving epidemiologic linkage. PNAS. 2016 Mar;113(10):2690–2695.

[39] Worby CJ, Lipsitch M, Hanage WP. Shared Genomic Variants: Identification of Transmission Routes Using Pathogen Deep-Sequence Data. Am J Epidemiol. 2017 Nov;186(10):1209–1216.

[40] Wymant C, Hall M, Ratmann O, Bonsall D, Golubchik T, de Cesare M, et al. PHYLOSCANNER: Inferring Transmission from Within- and Between-Host Pathogen Genetic Diversity. Mol Biol Evol. 2018;35(3):719–733.

[41] Volz EM, Siveroni I. Bayesian phylodynamic inference with complex models. PLOS Computational Biology. 2018 Nov;14(11):e1006546.

[42] Kühnert D, Kouyos R, Shirreff G, Pečerska J, Scherrer AU, Böni J, et al. Quantifying the fitness cost of HIV-1 drug resistance mutations through phylodynamics. PLOS Pathogens. 2018 Feb;14(2):e1006895.

[43] Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, et al. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLoS Computational Biology. 2014 Apr;10(4):e1003537.

[44] Gray RR, Parker J, Lemey P, Salemi M, Katzourakis A, Pybus OG. The mode and tempo of hepatitis C virus evolution within and among hosts. BMC Evol Biol. 2011;11:131.

[45] Wallinga J, Lipsitch M. How generation intervals shape the relationship between growth rates and reproductive numbers. Proc R Soc Lond B. 2007;274:599–604.

[46] Danesh G, Saulnier E, Gascuel O, Choisy M, Alizon S. Simulating trajectories and phylogenies from population dynamics models with TiPS. 2020;p. 1–7.

[47] Gillespie DT. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. Journal of Computational Physics. 1976;22(4):403–434.

[48] Lewitus E, Morlon H. Characterizing and comparing phylogenies from their laplacian spectrum. Systematic Biology. 2016;65(3):495–507.

[49] Csillery K, Francois O, Blum MGB. abc: an R package for approximate Bayesian computation (ABC). Methods in Ecology and Evolution. 2012;.

[50] Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society Series B (Statistical Methodology). 2005;67(2):301–320.